用户任务预测研究进展与算法分析

胡志明[1,2]， 李 胜[1,2]， 盖 孟[1,2]

(1. 北京大学信息科学技术学院，北京 100871；
2. 北京大学北京市虚拟仿真与可视化工程技术研究中心，北京 100871)

摘 要：用户在执行不同任务时，会表现出不同的感知行为。知道用户正在执行的任务可以帮助进行用户行为的分析，也可以作为智能交互系统的输入，使得系统自动根据用户不同的任务提供不同的功能，改善用户的体验。用户任务预测指的是根据用户的眼睛运动特征、场景内容特征等相关信息来预测用户正在执行的任务。用户任务预测是视觉研究领域中的一个热门研究课题，研究者们针对不同的场景提出了很多有效的任务预测算法。然而，以往工作中提出的算法大多是针对一种特定类型的场景，且不同算法之间缺乏统一的测试和分析。本文首先回顾了图片场景、视频场景、以及现实场景中用户任务预测问题的相关进展，接着对目前主要的任务预测算法进行了详细的介绍。并在一个现实场景任务数据集上对相关算法进行了测试和分析，为未来的相关研究提供了有意义的见解。

关 键 词：用户任务预测；感知状态预测；任务分类；扫描路径分类；机器学习

中图分类号：TP 391 **DOI**：10.11996/JG.j.2095-302X.2021030367

文献标识码：A **文 章 编 号**：2095-302X(2021)03-0367-09

Research progress of user task prediction and algorithm analysis

HU Zhi-ming[1,2], LI Sheng[1,2], GAI Meng[1,2]

(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;
2. Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University, Beijing 100871, China)

Abstract: Users' cognitive behaviors are dramatically influenced by the specific tasks assigned to them. Information on users' tasks can be applied to many areas, such as human behavior analysis and intelligent human-computer interfaces. It can be used as the input of intelligent systems and enable the systems to automatically adjust their functions according to different tasks. User task prediction refers to the prediction of users' tasks at hand based on the characteristics of his or her eye movements, the characteristics of scene content, and other related information. User task prediction is a popular research topic in vision research, and researchers have proposed many successful task prediction algorithms. However, the algorithms proposed in prior works mainly focus on a particular scene, and comparison and analysis are absent for these algorithms. This paper presented a review of prior works on task prediction in scenes of images, videos, and real world, and detailed existing task prediction algorithms. Based on a real-world task dataset, this paper evaluated the performances of existing algorithms and conducted the corresponding analysis and discussion. As such, this work can provide meaningful insights for future works on this important topic.

收稿日期：2021-03-15；**定稿日期**：2021-04-19

Received：15 March，2021；**Finalized**：19 April，2021

基金项目：国家自然科学基金项目(61632003)

Foundation items：National Natural Science Foundation of China (61632003)

第一作者：胡志明(1995–)，男，安徽安庆人，博士研究生。主要研究方向为人机交互与虚拟现实。E-mail：jimmyhu@pku.edu.cn

First author：HU Zhi-ming (1995–), male, PhD candidate. His main research interests cover human-computer interaction and virtual reality. E-mail：jimmyhu@pku.edu.cn

通信作者：盖 孟(1988–)，男，山东莱阳人，助理研究员，博士。主要研究方向为计算机图形学、虚拟仿真等。E-mail：gm@pku.org.cn

Corresponding author：GAI Meng (1988–), male, research associate, Ph.D. His main research interests cover computer graphics, virtual reality and simulation, etc. E-mail：gm@pku.org.cn

Keywords: user task prediction; cognitive state prediction; task classification; scanpath classification; machine learning

用户在执行不同的任务时，会表现出不同的感知行为[1-6]。知道用户正在执行的任务可以帮助研究者们更好地理解用户的行为。用户执行的任务这一信息也可以作为智能交互系统的输入[7-10]，让系统能够根据用户不同的任务来实现不同的功能，提升系统的智能化水平，改善用户的体验。

用户任务预测指的是根据用户的眼睛运动特征、场景内容特征等相关信息来预测用户正在执行的任务。用户任务预测是视觉研究领域中的一个热门研究课题，受到了研究者们极大的关注和重视。并针对图片场景、视频场景、以及现实场景开展了大量的相关工作，提出了很多有效的任务预测算法。

然而，以往工作中提出的算法往往都是针对某一种特定类型的场景，例如人物图片场景，这些算法在其他类型场景中的表现还有待研究。此外，不同的任务预测算法之间缺乏统一的测试和分析。

本文首先回顾了图片场景、视频场景以及现实场景中用户任务预测问题的研究进展，接着对目前主要的几种任务预测算法，即线性判别分析算法(linear discriminant analysis，LDA)、支持向量机(support vector machine，SVM)算法、Boosting 算法、随机森林算法(random forest，RFo)以及随机蕨算法(random ferns，RFe)进行了详细的介绍，并在一个现实场景任务数据集上对相关算法进行了测试和分析。本文的工作对未来有关用户任务预测问题的研究具有重要的指导意义。

1 用户任务预测研究进展

1.1 问题的提出

在 1967 年，YARBUS[11]针对用户执行的任务对其眼睛运动的影响开展了一项定性研究。其使用一张人物图片作为实验场景，收集了一名用户分别在 7 个任务下的眼睛运动数据。这 7 个任务分别是：①自由观看图片；②判断图片中家庭的物质条件；③判断图片中人物的年龄；④猜测客人到来前该家庭做了什么；⑤记住人物的衣服；⑥记住房间里的人物和物体的位置；⑦判断客人和这个家庭分离了多长时间。图 1 展示了文献[11]使用的实验图片以及用户在这 7 个任务下的眼睛运动数据。发现用户执行的任务对其眼睛的运动产生了极大的影响。

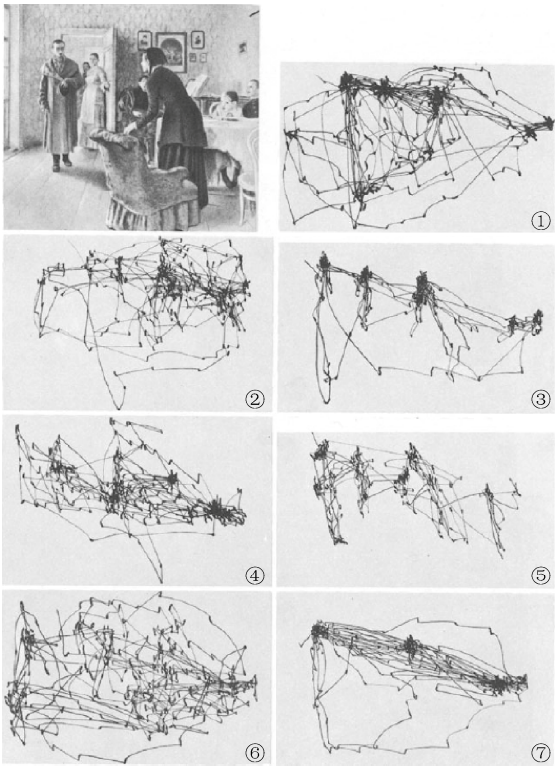

图 1 YARBUS[11]使用的实验图片及其记录的用户在 7 个任务下的眼睛运动数据

Fig. 1 The image used in YARBUS's experiment and the eye movement data in the corresponding seven tasks[11]

文献[11]的实验结果暗示了可以通过用户的眼睛运动来预测用户的任务。受到启发，大量研究者开始探索逆 YARBUS 问题(inverse YARBUS problem)，即用户任务预测问题，指的是通过用户的眼睛运动信息来预测用户正在执行的任务。YARBUS 问题研究的是视觉任务到眼睛运动的映射，而逆 YARBUS 问题(用户任务预测问题)研究的则是眼睛运动到视觉任务的映射。针对逆 YARBUS 问题，研究者们尝试了使用用户的眼睛运动特征、场景内容特征等相关信息来预测用户正在执行的任务，并且在图片场景、视频场景、以及现实场景中均取得了显著的进展。

1.2 图片场景任务预测

图片场景的用户任务预测问题是计算机视觉和感知科学中的一个热门研究课题，受到了研究者们极大的关注和重视。

GREENE 等[12]研究了人物图片场景中的用户任务预测问题。其收集了 64 张人物图片，每张图片至少包含 2 个人物。用户被要求去观看图片，并完成 4 个任务：记忆(memory)，记住图片的内容；年代(decade)，判断图片拍摄的年代；人物(people)，判断图片中人物之间彼此熟悉的程度；财富(wealth)，判断图片中人物的财富多少。其收集了用户分别在以上 4 个任务中的眼睛运动数据，用于研究该场景中的任务预测问题。

基于文献[12]的工作，KANAN 等[13]对用户的眼睛运动数据重新进行了分析，并使用了 SVM 算法来预测用户的任务，取得了更好的预测效果。

BORJI 和 ITTI[14]重新分析了文献[12]收集的数据，并进行了新的实验来收集数据。其使用了 15 张人物图片，并且要求用户在观看图片时，完成文献[11]原始的 7 个任务。用户的眼睛运动数据被记录下来，用于任务预测算法的训练和测试。采用 Boosting 算法来预测用户的任务，该算法在文献[12]收集的数据和新收集的数据上都取得了良好的预测效果。

KÜBLER 等[15]也开展了与文献[12]相似的工作。并使用 2 张人物油画作为实验场景，收集了用户在自由观察和年龄估计(估计油画中人物的年龄) 2 个不同任务下的眼睛运动数据。采用 SVM 算法来预测用户的任务，并且在新收集的数据、文献[12]和[14]的数据上分别进行了测试。结果表明，该算法具有较好的预测效果。

在文献[12]和[14]工作的基础上，FUHL 等[16]提出了一种随机蕨算法来预测用户的任务，并分别在文献[12]和[14]的数据上进行了模型的训练和测试。结果表明，该算法的效果明显优于之前的方法。

文献[7]关注自然图片和文本图片中的任务预测问题。其使用了 196 张自然图片和 140 张文本图片作为实验场景。其中，自然图片包含了室内和室外的环境；文本图片则取自网上的新闻报道，且包含了 40~60 个单词。用户在观看图片时，被要求完成 4 个不同的任务：场景记忆(scene memorization)，记忆场景的内容并完成相应的记忆测试；阅读(reading)，阅读文本的内容；场景搜索(scene search)，在场景图片中搜索嵌入的目标字母(“L”或“T”)；伪阅读(pseudo-reading)，阅读一些伪文本，伪文本中的文字由一些小方块组成。文献[7]收集了用户在这 4 种任务下的眼睛运动数据，且进行用户任务的预测。

KOEHLER 等[17]研究了自然图片场景中，不同的任务对用户视觉注意的影响。其使用了 800 张室内和室外的场景自然图片作为实验场景。用户被要求完成 3 个不同的任务即自由观察(free viewing)、显著性搜索(saliency search)以及特定目标搜索(cued object search)。并收集了用户在这 3 个任务中的眼睛运动数据，还分析了不同的任务对用户眼睛运动产生的影响。

基于文献[17]收集的数据，BOISVERT 和 BRUCE[18]研究了用户注视位置的空间分布、用户注视位置的动态信息以及用户观察的图片内容 3 方面特征在任务预测这一问题中的重要性。并提出了一个随机森林算法，结合了以上 3 方面特征来预测用户的任务，取得了良好的预测效果。

COUTROT 等[19]提出了一种基于线性判别分析的任务预测方法。其使用隐马尔可夫模型从用户的眼睛运动数据中提取特征，用于算法的训练和测试。在文献[17]收集的数据上进行了模型的测试，结果表明，该算法具有较高的预测精度。

1.3　视频场景任务预测

研究者们还探索过视频场景的用户任务预测问题。

HILD 等[20]专注于动态视频场景中的用户任务预测，并使用了如图 2 所示的动态视频来进行用户数据的收集。该视频是由一个固定视角的摄像机在街道上拍摄的，其时长为 4 min。视频中的动态信息主要包括走动的行人以及车辆的往来。用户在观看视频的时候，被指派了 4 个任务：探索(explore)，观看视频以熟悉视频的内容；观察(observe)，观察

图 2　文献[20]研究的动态视频场景

Fig. 2　The dynamic video scene used in reference [20]

视频并检测行人和车辆违反交通规则的情况；搜索(search)，在视频中寻找特定穿着的路人；追踪(track)，追踪位于场景内的并且距离自己最近的物体。文献[20]收集了用户在这 4 种任务下的眼睛运动数据，用于预测用户的任务。

HADNETT-HUNTER 等[21]则研究了虚拟场景中不同的任务对用户视觉注意带来的影响。使用了如图 3 所示的 3 种虚拟场景，从左到右依次为室内办公室场景(indoor office space)、郊区街道场景(suburban street)以及沙漠垃圾场场景(desert junkyard)。用户被要求在场景中分别完成自由观察(free viewing)、目标搜索(object search)、以及路径导航(path navigation) 3 种不同的任务。其收集了用户在 3 种任务下的眼睛运动数据，用于分析不同的任务对用户视觉注意产生的影响。

1.4 现实场景任务预测

研究者们针对现实场景任务预测这一问题，开展了很多研究工作。

BULLING 等[22]针对办公室环境(office environment)，预测了用户日常进行的 6 种任务。图 4 为文献[22]所研究的办公室场景以及场景中相应的 6 种日常任务。其任务包括：拷贝文本(copy)、阅读打印下来的文件(read)、手写做笔记(write)、观看视频(video)、浏览网页(browse)、以及没有具体任务的空闲状态(null)。文献[22]收集了用户在 6 种任务下的眼动电波图(electrooculography，EOG)数据，用于进行用户任务的预测。

文献[8]对日常生活场景的任务预测问题进行了探索。图 5 为文献[8]所研究的日常生活场景，以及场景中相应的 4 种日常任务。从左至右的任务分别是：社交(social)，和别人进行互动；感知(cognitive)，专注在某件事上；物理(physical)，进行物理上的运动；空间(spatial)，进行空间上的移动。文献[8]收集了用户在日常生活中 4 种任务下的眼睛运动数据，用于进行用户任务的预测。

文献[9]专注于阅读场景，研究了预测用户阅读的文档类型这一问题。图 6 为文献[9]所研究的阅读场景以及所研究的 5 种文档类型，从左至右依次是

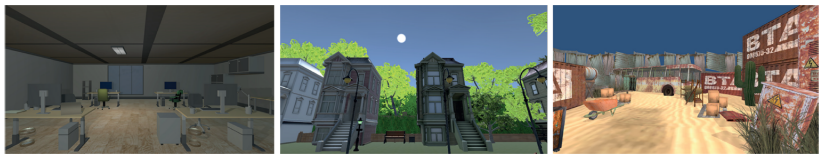

图 3 文献[21]使用的实验场景

Fig. 3 The experimental scenes used in reference [21]

图 4 文献[22]所研究的办公室场景和场景中相应的 6 种日常任务

Fig. 4 The office environment and the corresponding six tasks in reference [22]

图 5 文献[8]所研究的日常生活场景和场景中相应的 4 种日常任务

Fig. 5 The real-world environments and the corresponding four tasks in reference [8]

图 6 文献[9]研究的阅读场景和相应的 5 种文档类型

Fig. 6 The reading environments and the corresponding five document types in reference [9]

漫画书(manga comic)、课本(textbook)、时尚杂志(fashion magazine)、小说(novel)以及报纸(newspaper)。并收集了用户在阅读不同类型的文档时的眼睛运动数据，用于预测用户阅读的文档类型。

LIAO 等[23]研究了现实环境行人导航(pedestrian navigation)场景中的用户任务预测问题。5 种常见的导航任务(navigation task)分别是：定位自己的位置和方向(self-localization and orientation)、搜索局部环境中的目标(local environment target search)、搜索地图中的目标(map target search)、路线记忆(route memorization)以及步行至目的地(walking to the destination)。收集了用户在 5 种导航任务中的眼睛运动数据，用于预测用户的任务。

2　任务预测算法

2.1　线性判别分析

LDA 亦被称为 Fisher 判别分析，是一种经典的线性学习方法。该方法通过找到样本特征的一个线性组合，形成一个线性分类器，以用来区分不同类别的样本。

文献[20]将线性判别分析应用到了用户任务预测之中。首先采用 I-VT 算法[24]从原始的眼睛运动数据(raw gaze data)中提取了用户的注视(fixation)信息，并进一步提取了用户的注视特征来进行用户任务的分类。提取的注视特征包括注视持续时间(fixation duration)、眼跳幅度(saccade amplitude)、眼跳速度(saccade velocity)的均值和方差、每秒平均的注视数目(number of fixations per second)、注视直径(fixation diameter)的均值以及注视角度(fixation angle)的均值和方差。其中，注视直径是以属于一个注视的所有原始眼睛运动位置的最小包围圆(smallest enclosing circle)来计算的。注视角度则是由相邻 3 个注视的中心位置所形成的夹角来确定的。文献[20]以类内协方差(intra-class covariance)最小、并且类间协方差(inter-class covariance)最大为优化目标，学习不同注视特征的权重，并对特征进行线性组合形成一个线性判别分类器，区分用户不同的任务。

文献[19]也将线性判别分析应用到了任务预测之中。与文献[20]不同的是，其采用隐马尔可夫模型(hidden Markov models)从原始的眼睛运动数据中提取特征，并为用户的每一个扫描路径(scanpath)训练了一个隐马尔可夫模型，提取了 24 个特征值。接着学习了不同特征值的权重，以生成线性判别分类器，用于区分用户不同的任务。

2.2　支持向量机

SVM 是一种经典的机器学习分类模型。其基本思想是找到一个定义在特征空间上的间隔最大的线性分类器，对样本进行分类。通过引入核技巧，其将输入特征隐式映射到高维特征空间中，可以有效地实现非线性分类。

文献[22]将 SVM 应用到了用户任务预测的问题之中，采用眼动电波图测量技术(electrooculography，EOG)记录了用户在不同任务中的眼球运动信息，并从中提取了用户的注视(fixation)、眼跳(saccade)、以及眨眼(blink)等信息。其使用了极小冗余极大相关性(minimum redundancy maximum relevance)的特征提取方法，从原始的特征中选取了较为重要的特征，训练一个具有线性核函数的 SVM，并取得了良好的任务预测效果。

文献[8]也从用户的眼睛运动中提取了相应的特征进行 SVM 的训练，用于预测用户的任务。其将用户原始的眼睛运动编码为可以代表眼睛在不同方向运动的字符串(string of symbols)。其中，连续的眼睛运动被编码为具有不同长度的单词(word)，用来作为进行分类的基本特征。接着采用字符串核函数(string kernel function)将输入的字符特征映射到高维特征空间，以此进行 SVM 的学习和分类。

文献[13]也将 SVM 应用到了任务预测之中，并采用 Fisher 核学习(Fisher kernel learning)的方法[25]从原始的眼睛运动数据中提取 Fisher 核特征，还采用主成分分析(principal component analysis)的方法降低特征的维度。最后采用高斯径向基函数(Gaussian radial basis function)作为核函数，进行 SVM 的学习和分类。

COCO 和 KELLER[26]使用了用户眼睛运动的空间特征和时间特征来预测用户的任务。使用的特征包含了用户开始第一次眼睛运动的时间、用户注视的数目、眼跳幅度的均值、用户在物体上注视的总数以及场景中视觉注意空间分布的信息熵(the entropy of the attentional landscape)，并使用 SVM 对用户眼睛运动的空间和时间特征进行学习，用于预测用户的任务。

文献[15]从用户的眼睛运动数据中提取了序列特征来进行用户任务的预测，并将用户的扫描路径

(scanpath)编码为字符串，将其切分为多个短小的子序列。其提取了子序列的频率特征，训练一个具有线性核函数的 SVM，用于预测任务。

文献[19]则是使用了隐马尔可夫模型(hidden Markov models)从原始的眼睛运动数据中提取特征，接着将提取的特征输入到具有线性核函数的 SVM 中进行学习和预测。

2.3 Boosting 算法

Boosting 算法也称为提升算法，是一种经典的集成学习算法。其通过对训练样本的权重进行调整，学习多个不同的弱分类器，并进行线性组合，用以提升分类的效果。

文献[14]采用了 Boosting 算法中的 RUSBoost 进行用户任务的预测。RUSBoost 是一种针对数据类别不平衡问题的 Boosting 算法，其可通过随机欠采样(random under-sampling，RUS)的方式从训练数据集中抽取数据，用以进行弱分类器的训练。文献[14]从用户观察的图片内容以及用户的眼睛运动信息中提取了相应的特征，进行 RUSBoost 的训练和分类。提取的特征包括用户注视位置在图片内容上的分布图(fixation map)、归一化扫描路径显著性值(normalized scanpath saliency)的直方图、注视的数目、注视持续时间的均值、眼跳幅度的均值、注视区域覆盖整张图像的百分比以及用户的前 5 个注视位置。

文献[19]则使用了 Boosting 算法中的 AdaBoost 来进行用户任务的预测。其是一种非常具有代表性的 Boosting 算法，可通过提高前一轮错误分类样本的权重以及降低正确分类样本的权重，以此不断调整训练数据的权重来迭代地训练多个弱分类器。接着采用加权多数表决的方式，即加大分类误差小的弱分类器的权重、减小分类误差大的弱分类器的权重，并将多个弱分类器进行组合分类。文献[19]采用了隐马尔可夫模型提取用户眼睛运动数据中的特征，并进而用于 AdaBoost 的训练和测试。

2.4 随机森林

RFo 是一种简单、高效的集成学习算法。其以决策树(decision tree)作为基本的弱分类器，并且在决策树的训练过程中使用了随机属性选择的策略，使得集成后的分类器具有更好的泛化性能。

SUGANO 等[27]将 RFo 应用到了用户任务预测之中，并提取了多种不同的用户注视和眼跳特征，进行 RFo 的训练和测试。提取的用户注视特征包括注视位置的均值、方差、协方差(covariance)，注视持续时间的均值、方差、总和，注视起始时间的均值、方差以及注视的总数。提取的眼跳特征包括眼跳方向的均值、方差、协方差，眼跳长度和眼跳持续时间的均值、方差、总和，眼跳起始时间的均值、方差以及眼跳的总数。RFo 以注视和眼跳的特征作为输入，进行用户任务的预测。

文献[18]也使用了 RFo 来预测用户的任务，并从用户注视位置的空间分布、用户注视位置的动态信息以及用户观察的图片内容上提取了不同的特征，研究其与用户任务之间的联系。具体来说，其统计了用户注视位置在图片内容上的空间分布，并提取了用户注视位置的分布密度图(fixation density map)。共使用了包含 48 个滤波器的 Leung–Malik 滤波器组(Leung–Malik filter bank)[28]，从图片内容上用户注视位置所在的区域上提取了相应的图像特征。也从图片中用户注视的区域中计算了不同方向梯度的直方图分布(histogram of oriented gradients)。还使用 Gist 描述子[29]从图像内容中提取了场景的整体结构特征。最后，将提取的所有特征整合到一起，进行 RFo 的训练，并取得了良好的预测效果。

文献[20]也将 RFo 应用到了用户任务预测这一问题之中，并从用户的注视信息中提取了相应的特征，用于进行 RFo 的训练。提取的特征包括注视持续时间的均值和方差、眼跳幅度的均值和方差、眼跳速度的均值和方差、每秒平均的注视数目、注视直径的均值以及注视角度的均值和方差。RFo 由一系列的决策树组成，文献[20]尝试了不同数目的决策树组合，最终选定了 100 个决策树组成了 RFo，用于预测用户的任务。

文献[19]也使用了 RFo 来预测用户的任务。利用了隐马尔可夫模型从原始的眼睛运动数据中提取了相应的特征，进行 RFo 的训练和测试。

文献[23]从用户的眼睛运动数据中提取了统计特征、空间特征以及时间特征，用于进行 RFo 的学习。具体来说，眼睛运动的统计特征包括注视、眼跳、眨眼以及瞳孔直径(pupil diameter) 4 个典型眼睛运动参数的基本统计信息，例如频率、最大值、最小值、均值及偏度(skewness)。眼睛运动的空间特征包括注视分布特征和眼跳方向特征。眼睛运动的时间特征包括不同时间分段(time slicing)的统计特征，以及从眼跳时间序列中提取的特征。所有的特征都被作为 RFo 的输入，用于模型的训练和预测。

2.5　随机蕨

RFe 是一种以蕨(fern)算法作为基本分类器的集成学习算法，具有容易训练、分类速度快、分类精度高等优点。蕨算法通过对输入特征进行 S (S 是蕨算法的尺寸参数)次二进制测试(binary test)，将输入特征映射为了一个长度为 S，每一位的值为 0 或者 1 的特征向量。将该特征向量转换到 10 进制，就得到了范围在$[0, 2^S–1]$之间的一个数值。换言之，蕨算法的功能是将输入特征映射为$[0, 2^S–1]$范围内的一个特征值。在训练的过程中，蕨算法将所有的输入特征都映射为特征值，并统计了属于每个类别的特征值直方图分布。在预测时，蕨算法先将输入特征映射为特征值，再从每个类别的直方图上查看该特征值的分布概率，并选取使得特征值具有最大分布概率的类别作为预测的类别。RFe 算法则是集成了多个蕨算法来进行预测。使用了多个独立的蕨算法，每个蕨算法随机选取了输入特征的一个子集作为输入来进行训练。对输入特征进行分类时，首先查找该输入特征的相应子集在各个蕨算法中得到的不同类别的分布概率，再将不同蕨算法得到的分布概率相乘，得到该输入特征在所有类别上的概率，最后选取概率最大的类别作为预测的类别。

文献[16]将 RFe 算法应用到了用户任务预测之中，并取得了良好的预测效果。其从用户原始的眼睛运动数据中提取了用户的注视和眼跳信息，以连续的眼跳角度(saccade angle succession)作为输入特征，进行 RFe 算法的训练和测试。分别在 2 个用户任务预测的数据集[12,14]上测试了模型的效果。结果表明，该算法具有较好的预测精度。

3　算法测试与分析

3.1　算法实现与评价指标

本文对第 2 节介绍的任务预测算法进行了测试。采用了文献[19]提供的 MATLAB 工具箱中实现的 LDA，SVM，Boosting 算法以及 RFo 来进行测试。该工具箱使用隐马尔可夫模型从用户原始的眼睛运动数据中提取了相应的特征，作为任务预测算法的输入进行模型的训练。本文采用了文献[16]提供的源代码，对 RFe 算法进行测试。

本文采用分类准确率，即正确分类的样本数占总样本数的比例，作为任务预测算法的评价指标。

3.2　测试数据集

本文采用最近发布的一个现实场景任务数据集，即 GW 数据集[30]进行算法的测试。该数据集收集了 19 名用户在现实环境中，执行 4 种不同任务时的眼睛运动数据，每个任务的持续时间是 3 min 左右。图 7 为 GW 数据集的实验场景和相应的 4 种任务。从左至右依次是：室内导航(indoor navigation)，在室内按照指定的路径来行走；接球(ball catching)，接住扔过来的球；视觉搜索(visual search)，在场景中搜索具有几何形状(例如三角形、矩形)的物体；沏茶(tea making)。

图 7　GW 数据集的实验场景和相应的 4 种任务[30]

Fig. 7　The experimental environment of GW dataset and the corresponding four tasks[30]

该数据集是目前公布的最大的任务数据集。因而，本文选择在该数据集上进行任务预测算法的测试。

3.3　测试结果

本文将 GW 数据集中的用户眼睛运动数据以 25 Hz 的频率进行采样，并以 10 s 为一个窗口来进行切片，并用于模型的训练。相邻 2 个窗口的间距设置为 1 s。本文采用五折交叉验证的方法来测试各个算法的效果。具体而言，本文将 GW 数据集中的数据按照用户的不同，平均分成了 5 份(五折)，每次使用其中 4 份数据进行模型的训练，在剩下的 1 份数据上进行测试，一共进行了 5 组测试。表 1 为五折交叉验证的测试结果。可以看到，RFo 算法在五折平均的预测表现是最优的。

表 1　不同任务预测算法的五折交叉验证测试结果(%)

Table 1　The five-fold cross validation results of different task prediction algorithms (%)

组别	LDA	SVM	Boosting	RFo	RFe
1	37.1	33.2	32.9	38.5	33.1
2	39.7	40.7	32.8	48.0	38.5
3	39.3	35.8	30.7	38.6	36.7
4	41.8	44.3	41.5	47.7	38.0
5	37.2	36.4	33.0	39.9	34.8
平均值	39.0	38.1	34.2	42.5	36.2

本文在一台处理器为 Intel(R) Xeon(R) CPU E3-1230 v5 @3.40 GHz，内存为 16.0 GB 的机器上，对各个算法的运行效率进行了测试。本文在测试时发现，各算法在训练时比较耗时，模型训练好后，测试时的效率都非常高，各算法对单个样本的测试时间均小于 1 ms。因而，本文着重测试了各算法的训练时间。表 2 为各算法的训练时间。可以看到，LDA 算法的训练效率显著高于其他算法。

表 2 不同任务预测算法的训练时间对比

Table 2 The training times of different task prediction algorithms

组别	LDA (s)	SVM (s)	Boosting (s)	RFo (s)	RFe (h)
1	0.01	69.9	2.3	6.7	2.4
2	0.01	57.9	2.4	7.5	2.3
3	0.01	68.2	2.3	7.0	2.2
4	0.01	49.9	2.4	7.3	2.4
5	0.01	60.7	2.5	7.8	2.5
平均值	0.01	61.3	2.4	7.3	2.4

3.4 分析与讨论

表 3 总结了本文测试的不同任务预测算法的特点。就预测精度而言，预测效果最好的是 RFo 算法，其次是 LDA 算法。就时间复杂度而言，LDA 算法的时间复杂度最低，Boosting 和 RFo 算法的时间复杂度也不高。综合预测精度和时间复杂度 2 方面来考虑，最好的 2 个任务预测算法分别是 LDA 和 RFo 算法。

表 3 不同任务预测算法的特点对比

Table 3 The characteristics of different task prediction algorithms

算法	输入特征	预测精度	时间复杂度
LDA	隐马尔可夫特征	较高	低
SVM	隐马尔可夫特征	一般	较高
Boosting	隐马尔可夫特征	低	较低
RFo	隐马尔可夫特征	高	一般
RFe	原始眼睛运动数据	较低	高

LDA 算法由于只使用了样本特征的一个线性组合生成一个线性分类器来进行分类，因而其时间复杂度非常低，训练效率特别高。RFo 算法在决策树的训练过程中使用了随机属性选择的策略，使得集成后的分类器具有更好的泛化性能，因而具有非常好的预测精度。

针对一般的用户任务预测问题，推荐先使用训练速度快的 LDA 算法获取一个初步的结果，再尝试使用 RFo 算法，取得更好的预测性能。

3.5 未来展望

目前，针对用户任务预测这一问题，研究者们大都只关注某一种特定类型的场景，例如自然图片场景、视频场景，并通过收集用户在场景中执行不同任务时的实验数据来进行用户任务的预测。由于一种特定类型的场景中收集的用户数据往往规模较小，研究者们通常都是采用所需训练样本较少的传统机器学习算法，例如 RFo 算法来进行用户任务的预测。

近年来，基于大量训练数据的深度学习算法开始被广泛地应用于各个领域之中，并且取得了很多突破性的成果。相信随着任务数据集规模的进一步扩大，深度学习算法很快也将被引入到用户任务预测这一问题之中，以实现更高的预测精度。

4 总 结

用户任务预测是视觉研究领域中的一个热门课题，任务预测算法在智能交互系统以及相关领域中具有重要的应用前景。本文重点回顾了图片场景、视频场景以及现实场景中用户任务预测问题的相关研究进展。本文深入介绍了目前主要的几种任务预测算法，即 LDA，SVM，Boosting 算法、RFo 和 RFe。本文进一步在一个现实场景任务数据集上测试了各个算法的效果，并进行了相关的分析与讨论。本文的工作对未来有关用户任务预测的研究具有重要的指导作用。

参考文献 (References)

[1] HU Z M, ZHANG C Y, LI S, et al. SGaze: a data-driven eye-head coordination model for realtime gaze prediction[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(5): 2002-2010.

[2] HU Z M. Gaze analysis and prediction in virtual reality[C]// 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). New York: IEEE Press, 2020: 543-544.

[3] HU Z M, LI S, GAI M. Temporal continuity of visual attention for future gaze prediction in immersive virtual reality[J]. Virtual Reality & Intelligent Hardware, 2020, 2(2): 1-11.

[4] HU Z M, LI S, ZHANG C Y, et al. DGaze: CNN-based gaze prediction in dynamic scenes[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(5): 1902-1911.

[5] HU Z M, BULLING A, LI S, et al. FixationNet: forecasting eye fixations in task-oriented virtual environments[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(5): 2681-2690.

[6] HU Z M. Eye Fixation Forecasting in Task-Oriented Virtual Reality[C]//2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). New York: IEEE Press, 2021: 707-708.

[7] HENDERSON J M, SHINKAREVA S V, WANG J, et al. Predicting cognitive state from eye movements[J]. PLoS ONE, 2013, 8(5): e64937.

[8] BULLING A, WEICHEL C, GELLERSEN H. EyeContext: recognition of high-level contextual cues from human visual behaviour[C]//2013 SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2013: 305-308.

[9] KUNZE K, UTSUMI Y, SHIGA Y, et al. I know what you are reading: recognition of document types using mobile eye tracking[C]//2013 International Symposium on Wearable Computers. New York: ACM Press, 2013: 113-116.

[10] LETHAUS F, BAUMANN M R, KöSTER F, et al. A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data[J]. Neurocomputing, 2013, 121: 108-130.

[11] YARBUS A L. Eye movements and vision[M]. Heidelberg: Springer, 1967: 171-211.

[12] GREENE M R, LIU T, WOLFE J M. Reconsidering Yarbus: a failure to predict observers' task from eye movement patterns[J]. Vision Research, 2012, 62: 1-8.

[13] KANAN C, RAY N A, BSEISO D N, et al. Predicting an observer's task using multi-fixation pattern analysis[C]//2014 Symposium on Eye Tracking Research & Applications. New York: ACM Press, 2014: 287-290.

[14] BORJI A, ITTI L. Defending Yarbus: eye movements reveal observers' task[J]. Journal of Vision, 2014, 14(3): 29-29.

[15] KÜBLER T C, ROTHE C, SCHIEFER U, et al. SubsMatch 2.0: scanpath comparison and classification based on subsequence frequencies[J]. Behavior Research Methods, 2017, 49(3): 1048-1064.

[16] FUHL W, CASTNER N, KÜBLER T, et al. Ferns for area of interest free scanpath classification[C]//The 11th ACM Symposium on Eye Tracking Research & Applications. New York: ACM Press, 2019: 1-5.

[17] KOEHLER K, GUO F, ZHANG S, et al. What do saliency models predict?[J]. Journal of Vision, 2014, 14(3): 14-14.

[18] BOISVERT J F, BRUCE N D. Predicting task from eye movements: on the importance of spatial distribution, dynamics, and image features[J]. Neurocomputing, 2016, 207: 653-668.

[19] COUTROT A, HSIAO J H, CHAN A B. Scanpath modeling and classification with hidden Markov models[J]. Behavior Research Methods, 2018, 50(1): 362-379.

[20] HILD J, VOIT M, KüHNLE C, et al. Predicting observer's task from eye movement patterns during motion image analysis[C]//2018 Symposium on Eye Tracking Research & Applications. New York: ACM Press, 2018: 1-5.

[21] HADNETT-HUNTER J, NICOLAOU G, O'NEILL E, et al. The effect of task on visual attention in interactive virtual environments[J]. ACM Transactions on Applied Perception (TAP), 2019, 16(3): 1-17.

[22] BULLING A, WARD J A, GELLERSEN H, et al. Eye movement analysis for activity recognition using electrooculography[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(4): 741-753.

[23] LIAO H, DONG W H, HUANG H S, et al. Inferring user tasks in pedestrian navigation from eye movement data in real-world environments[J]. International Journal of Geographical Information Science, 2019, 33(4): 739-763.

[24] SALVUCCI D D, GOLDBERG J H. Identifying fixations and saccades in eye-tracking protocols[C]//2000 Symposium on Eye Tracking Research & Applications. New York: ACM Press, 2000: 71-78.

[25] VAN DER MAATEN L. Learning discriminative fisher kernels[C]//2011 International Conference on Machine Learning. Washington, DC: Omnipress, 2011: 217-224.

[26] COCO M I, KELLER F. Classification of visual and linguistic tasks using eye-movement features[J]. Journal of Vision, 2014, 14(3): 11-11.

[27] SUGANO Y, OZAKI Y, KASAI H, et al. Image preference estimation with a data-driven approach: a comparative study between gaze and image features[J]. Journal of Eye Movement Research, 2014, 7(3): 5, 1-9.

[28] LEUNG T, MALIK J. Representing and recognizing the visual appearance of materials using three-dimensional textons[J]. International Journal of Computer Vision, 2001, 43(1): 29-44.

[29] OLIVA A, TORRALBA A. Building the gist of a scene: the role of global image features in recognition[J]. Progress in Brain Research, 2006, 155: 23-36.

[30] KOTHARI R, YANG Z, KANAN C, et al. Gaze-in-wild: a dataset for studying eye and head coordination in everyday activities[J]. Scientific Reports, 2020, 10(1): 1-18.